Table of Contents

Exploratory Data Analysis

Project Description: Exploratory Data Analysis (EDA) on DAP processing and summarizing bulk datasets required by the City of Vancouver

Project Title: AWS Data Analytic Platform for The City of Vancouver

Objective: The primary goal of this project is to perform exploratory data analysis (EDA) on the 311 contact center metrics dataset to design and implement a data analytic platform to support descriptive analysis. By analyzing various features such as Date, Calls_Offered, Calls_Handled, Calls_Abandoned, Average_Speed_of_Answer, Service_Level, and BI_ID, we aim to understand the factors that influenced the likelihood of the service of 311.

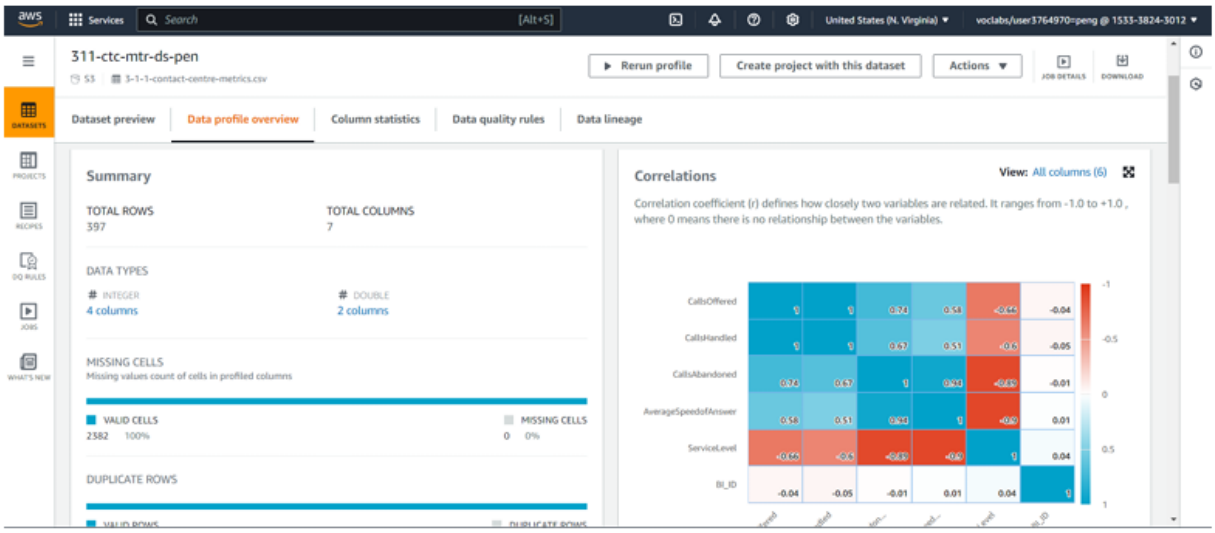
Dataset: This dataset consists of 311 service information, including details such as:

- Date: 2024-1-1 to 2025-1-31
- Calls_Offered: the time of 311 offering calls (seconds)
- Calls_Handled: the time of 311 handling calls (seconds)
- Calls_Abandoned: the time of 311 abandoning calls (seconds)
- Average_Speed_of_Answer: the time of average speed of answers from 311 (seconds)
- Service_Level: 311 service level (0-1)
- BI_ID: The ID number of providing services

When uploading the dataset in AWS, the data profile overview (see Figure 1) can show the summary of this dataset.

Figure 1: Data Profile Overview

Data Profile Overview



Note. The screenshot shows the data profile overview. Source from AWS.

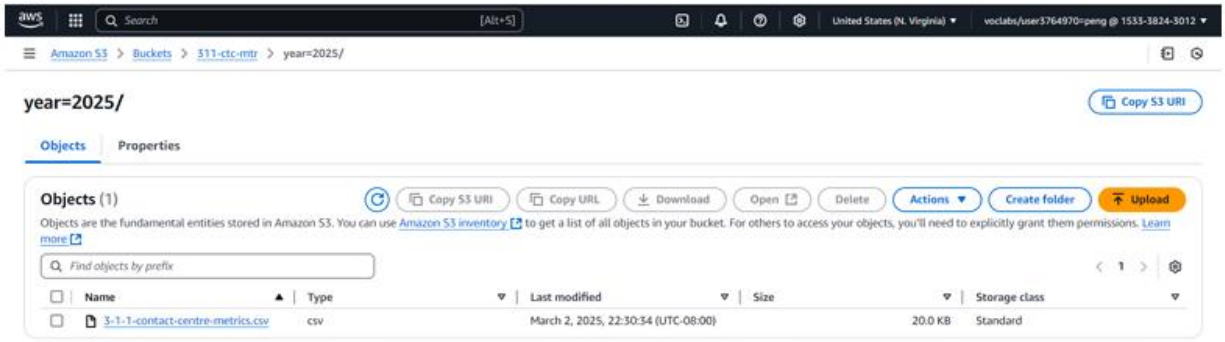
Methodology:

1. Data Collection and Preparation and Cleaning:

- Download the dataset from the City of Vancouver
- Observing the initial dataset, which includes missing values, outliers, and data type.
- The initial data is mostly cleaning, so I just adjusted the two decimal places and then changed this dataset to a CSV file to be uploaded to AWS S3 Buckets (see figure 2 and figure 3)

Figure 2: S3 Bucket Building

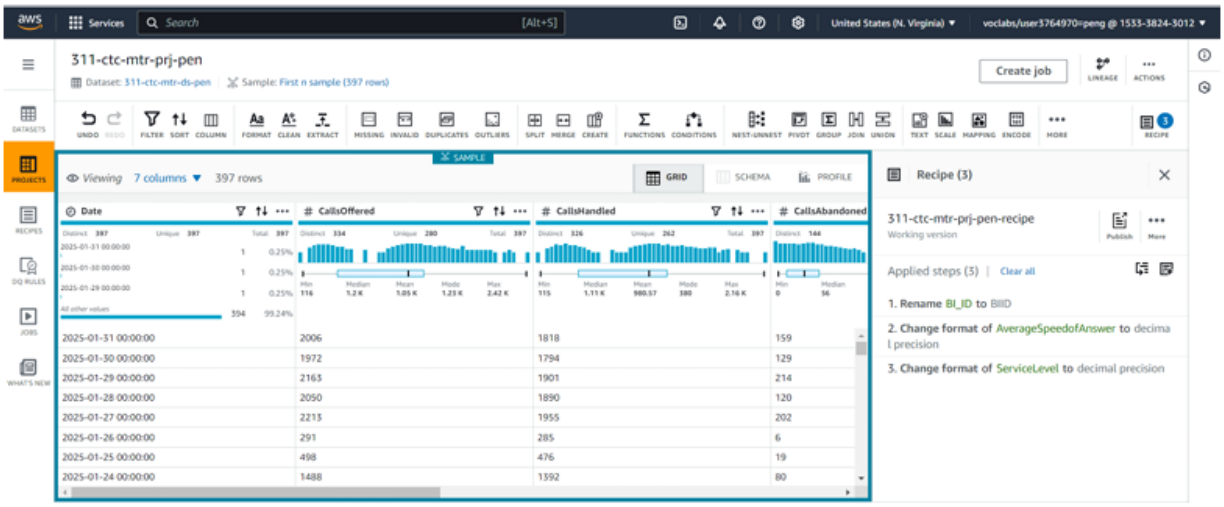
S3 Bucket Building



Note. The screenshot shows the S3 bucket. Source from AWS.

Figure 3: Data Cleaning

Data Cleaning



Note. The screenshot shows the data cleaning. Source from AWS.

2. Descriptive Statistics:

All data are numerical values, but Service_Level and Average_Speed_of_Answer are not integers.

3. Data Cataloging:

I created a new bucket called 311-ctc-mtr-cur and sub-folders for metrics including system and user folders, which will be stored for summarization. When the Crawler was ready, the cleaning data was transferred to databases for 34 cataloging and prepared for summarization (See Figure 4, Figure 5, Figure 6).

Figure 4: Metrics Folders Building

Metrics Folders Building

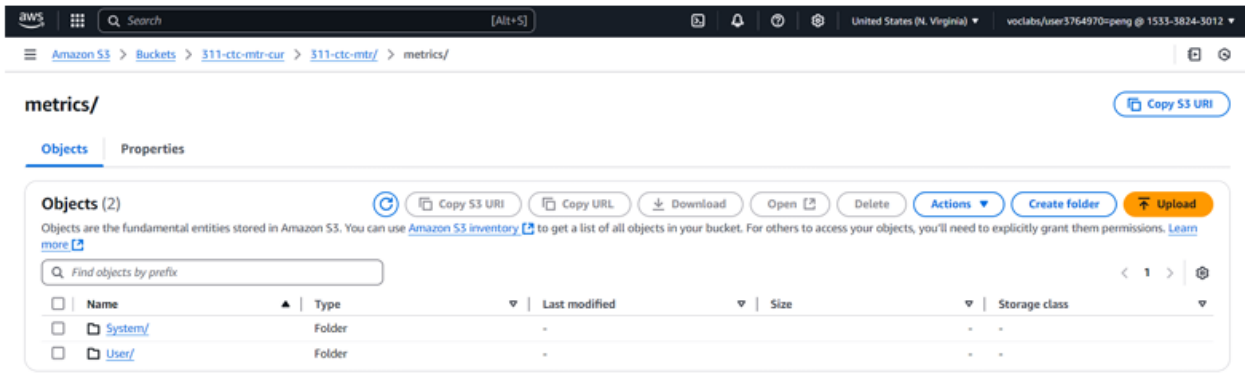
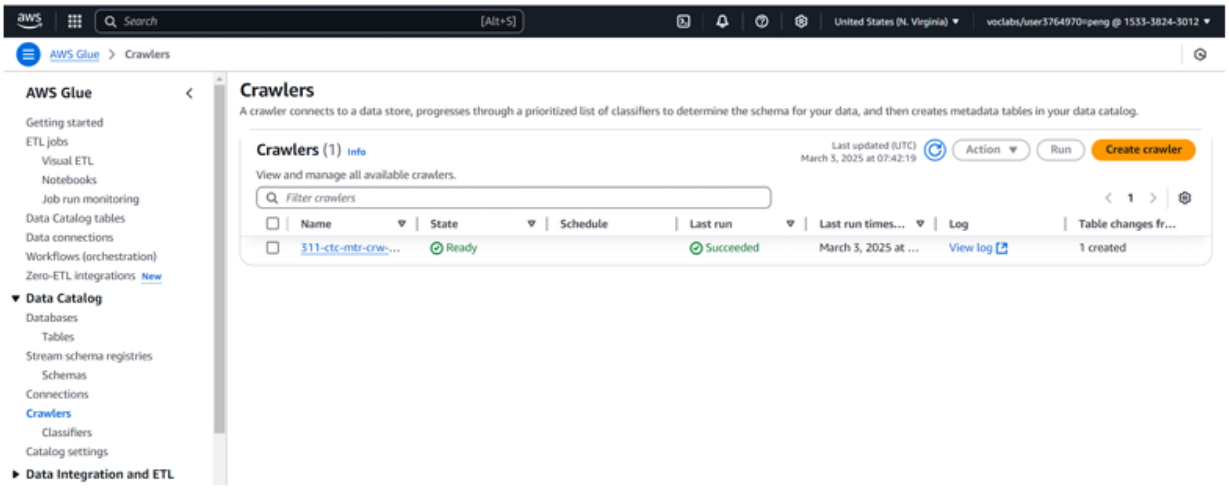


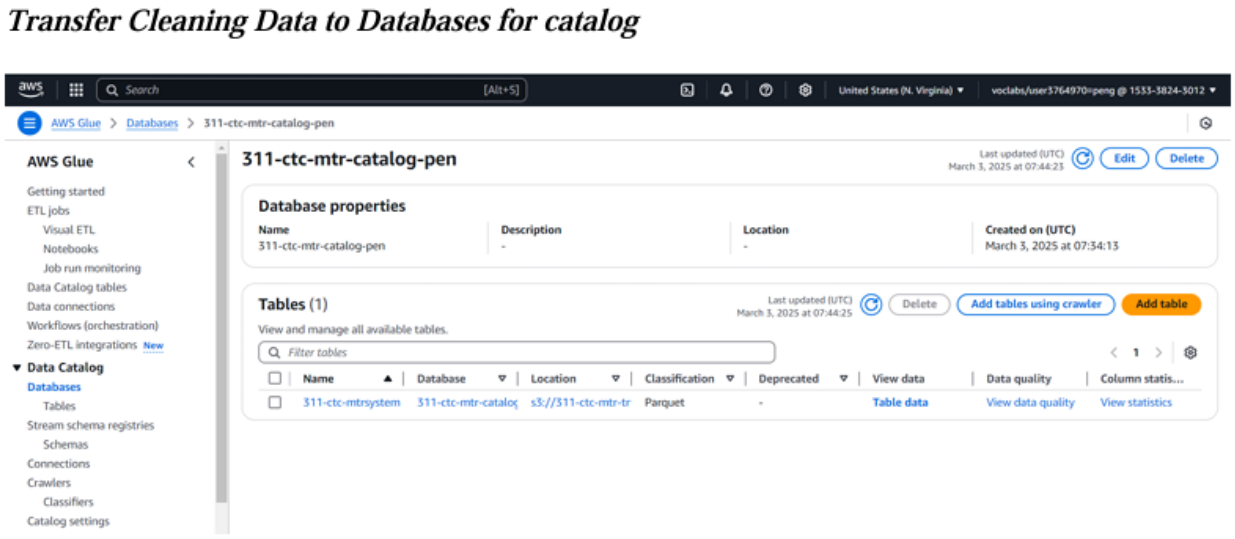
Figure 5: Crawlers Building

Crawlers Building



Note. The screenshot shows the crawler building. Source from AWS.

Figure 6: Transfer Cleaning Data to Databases for Catalog



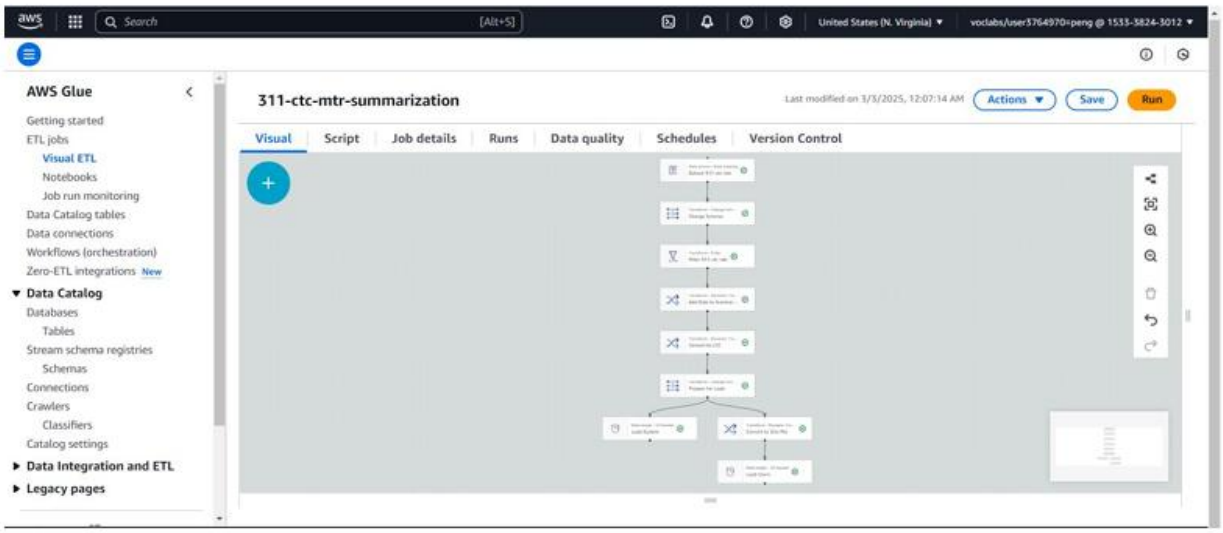
Note. The screenshot shows the cleaning data has already been transferred to datasets for catalog. Source from AWS.

4. Data Summarization:

Two ETL pipelines were established, and the metrics report was for systems and users.

Figure 7: ETL Pipeline

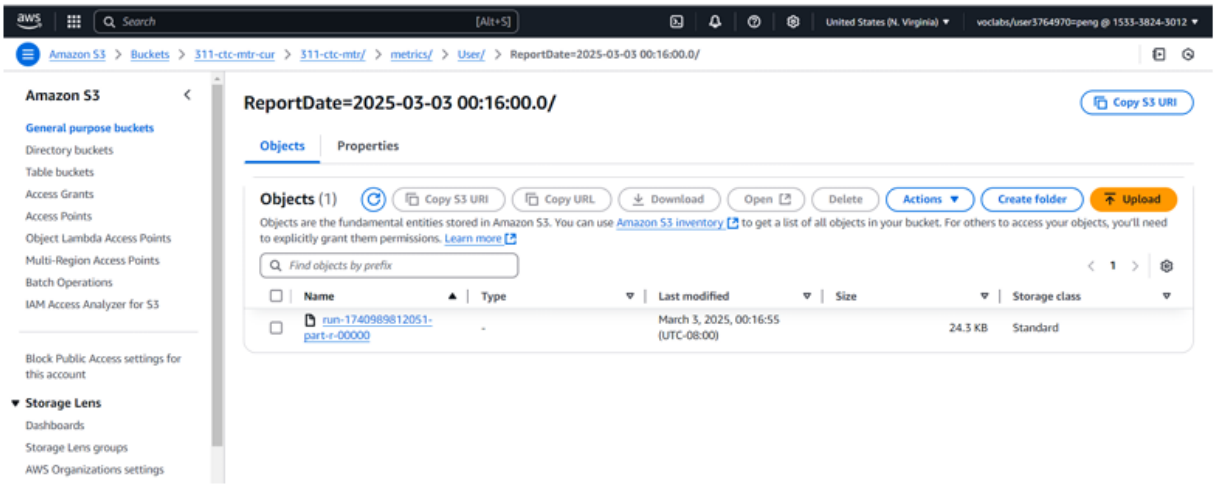
ETL pipeline 1



Note. The screenshot shows dataset summarization. Source from AWS.

Figure 8: Metrics Report

Metrics Report for User



Note. The screenshot shows the metrics report for users. Source from AWS.

5. Conclusion:

This project designed and implemented a data analytic platform that supports descriptive analysis and provided a thorough report that breaks down each step of the process. City of Vancouver can leverage AWS to drive data-based decision-making with minimal cost and high efficiency

6. Data Cleaning (in-class activity):

Table 1: Poor Data Quality Dataset

ID	Name	Age	Date of Birth	Country	Income	Gender	Transaction Date	Height (cm)	Satisfaction
1	Alice Johnson	30	1/2/1993	USA	$50,000	F	2/5/2023	170	7
2	Bob Smith	29	5/12/1994	United States	55000	Male	5/2/2023	180	8
3	Charlie Brown	-5	10/10/2015	US	$100,000	M	2/1/2023		6
4	Dana Lee	45	11/11/1978	USA	â‚¬60,000	female	13-02-2023	165	9
5	Eve Adams		7/25/1985	USA	$70,000	F	2/15/2023	172	
6	Frank Miller	35	12/12/1988		$120,000	M	2023.02.10	175	10
7	Alice Johnson	30	1/2/1993	USA	$50,000	F	2/5/2023	170	7
	duplicate	missing	inconsistent format	missing	inconsistent format	inconsistent format	inconsistent format	missing	missing
		invalid	unbalanced		typo				
					unbalanced				
					encoding				

In this class activity, we practiced identifying the poor data quality including duplicate, missing values, invalid values, inconsistent format, unbalanced, typo, encoding, and so forth.

Tools and Technologies:

AWS S3, Glue, Glue DataBrew

Descriptive Analysis

Project Description: Descriptive Analysis of Sales Performance (W6 in-class)

Project Title: Understanding the Sales Performance

Objective: Questions about Sales Performance Analysis, Impact of Marketing Spend, Effect of Temperature on Sales, and Website Traffic Insight were answered.

Dataset: The dataset includes the following key features:

- Date: 2023-1-1 to 2024-3-31

- Sales: Sales of each corresponding date
- Temperature: the temperature of each corresponding date
- Marketing Spend: the expenditure on the marketing of each corresponding date
- Customer purchase behavior: there are three categories, including A,B, and C.
- Website Traffic: the amount of traffic of each corresponding date
- Outlier Flag: there are two

Table 2: Sales Performance Dataset

Date	Sales	Temperature	Marketing_Spend	Customer_purchase_behavior	Website_Traffic	Outlier_Flag
1/1/2023	80	5	100	A	500	0
1/1/2023	70	5	100	A	500	0
1/2/2023	60	6	120	A	1000	0
1/2/2023	100	6	100	A	20	0
1/15/2023	100	8	100	B	700	0
1/15/2023	50	8	200	B	800	0
1/31/2023	50	12	100	B	800	0
1/31/2023	200	12	300	B	1000	0
2/14/2023	70	10	220	C	900	0
2/14/2023	200	10	200	C	1000	0
2/29/2023	120	15	100	C	2000	1
2/29/2023	200	15	400	C	2500	1
3/15/2023	300	18	100	A	1500	0
3/15/2023	100	18	500	A	1500	0
3/31/2023	200	22	400	A	500	0
3/31/2023	300	22	300	A	100	0
1/1/2024	70	5	100	A	500	0
1/1/2024	80	5	100	A	500	0
1/2/2024	100	6	100	A	20	0
1/2/2024	60	6	120	A	1000	0
1/15/2024	50	8	200	B	800	0
1/15/2024	150	8	100	B	700	0
1/31/2024	200	12	300	B	1000	0
1/31/2024	50	12	100	B	800	0
2/14/2024	200	10	200	C	1000	0

Methodology:

1. Data Collection and Preparation:

- Load the dataset using data analysis tools, including Excel and SQL
- The original data is cleaned.

2.1 Descriptive Statistics:

Table 3: Questions about Sales Performance Analysis, SQL Syntax, and Analysis

1. Sales Performance Analysis				
What is the **average sales** over the given period?	Distribution	SELECT AVG(Sales) AS Average_Sales FROM sales_table;	**Average_Sales**	
			139.0625	
What is the **highest and lowest sales** recorded?	Distribution	SELECT MAX(Sales) AS Highest_Sales, MIN(Sales) AS Lowest_Sales FROM sales_table;	**Highest_Sales**	**Lowest_Sales**
			300	50
How does sales vary **by date**?	Trend (Time-Series)	SELECT Date, SUM(Sales) AS Total_Sales FROM sales_table GROUP BY Date ORDER BY Date;	**Date**	**Total_Sales**
			1/1/2023	150
			1/1/2024	150
			1/15/2023	150
			1/15/2024	200
			1/2/2023	160
			1/2/2024	160
			1/31/2023	250
			1/31/2024	250
			2/14/2023	270
			2/14/2024	270
			2/29/2023	320
			2/29/2024	320
			3/15/2023	400
			3/15/2024	400
			3/31/2023	500

Calculate summary statistics for key variables, including:

- Average sales
- The highest and lowest sales
- How do sales vary by date

3.1 Data Visualization:

Table 4: Data Visualization by Excel

Average_Sales	
139.0625	
Highest_Sales	**Lowest_Sales**
300	50
Date	**Total_Sales**
1/1/2023	150
1/1/2024	150
1/15/2023	150
1/15/2024	200
1/2/2023	160
1/2/2024	160
1/31/2023	250
1/31/2024	250
2/14/2023	270
2/14/2024	270
2/29/2023	320
2/29/2024	320
3/15/2023	400
3/15/2024	400
3/31/2023	500

Create visual representations to illustrate findings:

- Average sales: 139.0625
- The highest sales: 300
- The lowest sales: 50
- Sales vary in each date

2.2 Descriptive Statistics:

Table 5: Questions about Impact of Marketing Spend, SQL Syntax, and Analysis

2. Impact of Marketing Spend				
What is the **correlation** between **marketing spend** and **sales**?	Correlation (Cause-Effect)	SELECT CORR(Marketing_Spend, Sales) AS Correlation_Marketing_Sales FROM sales_table;	Correlation_Marketing_Sales	
			0.313634298	
Does **higher marketing spend** result in **higher website traffic**?	Correlation (Cause-Effect)	SELECT CORR(Marketing_Spend, Website_Traffic) AS Correlation_Marketing_Website FROM sales_table;	Correlation_Marketing_Website	
			0.318149461	
Which dates had the **highest and lowest marketing spend**?	Comparison	SELECT Date, Marketing_Spend FROM sales_table ORDER BY Marketing_Spend DESC LIMIT 1 UNION SELECT Date, Marketing_Spend FROM sales_table ORDER BY Marketing_Spend ASC LIMIT 1;	Date	Marketing_Spend
			3/15/2023	500
			1/1/2023	100

Calculate summary statistics for key variables, including:

- Correlation between marketing spend and sales
- The correlation between marketing spend and website traffic
- Which dates had the highest and lowest marking spending

3.2 Data Visualization:

Table 6: Data Visualization by Excel

Correlation_Marketing_Sales
0.313634298
Correlation_Marketing_Website
0.318149461

Date	Marketing_Spend
3/15/2023	500
1/1/2023	100

Create visual representations to illustrate findings:

- Correlation between marketing spend and sales: 0.313434298
- Correlation between marketing spend and website: 0.318149461
- The highest marketing spend date: 2023-3-15
- The lowest marketing spend date: 2023-1-1

2.3 Descriptive Statistics:

Table 7: Questions about Effect of Temperature on Sales, SQL Syntax, and Analysis

3. Effect of Temperature on Sales				
Is there a **relationship between temperature and sales**?	Correlation (Cause-Effect)	SELECT CORR(Temperature, Sales) AS Correlation_Temperature_Sales FROM sales_table;	Correlation_Temperature_Sales	
			0.701793526	
Does a **higher temperature lead to more or fewer sales**?	Trend (Time-Series)	SELECT Temperature, AVG(Sales) AS Avg_Sales FROM sales_table GROUP BY Temperature ORDER BY Temperature;	Temperature	Avg_Sales
			5	75
			6	80
			8	87.5
			10	135
			12	125
			15	160
			18	200
			22	250

Calculate summary statistics for key variables, including:

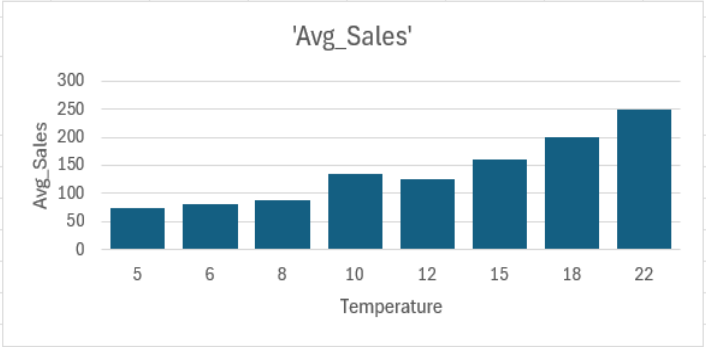
- Correlation between temperature and sales
- Does higher temperatures lead to more or fewer sales?

3.3 Data Visualization:

Table 8: Data Visualization by Excel

Correlation_Temperature_Sales
0.701793526

Temperature	Avg_Sales
5	75
6	80
8	87.5
10	135
12	125
15	160
18	200
22	250



Create visual representations to illustrate findings:

- Correlation between temperature and sales: 0.701793526
- When the temperature is higher, the average sales are higher.

2.4 Descriptive Statistics:

Table 9: Questions about website traffic insights, SQL Syntax, and Analysis

4. Website Traffic Insights				
What is the **average daily website traffic**?	Distribution	SELECT AVG(Website_Traffic) AS Average_Website_Traffic FROM sales_table;	Average_Website_Traffic	
			985.625	
On which dates was the **website traffic highest and lowest**?	Trend (Time-Series)	SELECT Date, Website_Traffic FROM sales_table ORDER BY Website_Traffic DESC LIMIT 1 UNION SELECT Date, Website_Traffic FROM sales_table ORDER BY Website_Traffic ASC LIMIT 1;	Date	Website_Traffic
			2/29/2023	2500
			44928	20
Does an **increase in website traffic** lead to **higher sales**?	Correlation (Cause-Effect)	SELECT CORR(Website_Traffic, Sales) AS Correlation_Website_Sales FROM sales_table;	Correlation_Website_Sales	
			0.224236537	

Calculate summary statistics for key variables, including:

- What is the average daily website traffic?
- On which dates was the website traffic highest and lowest?
- Does an increase in website traffic lead to higher sales?

3.4 Data Visualization:

Table 10: Data Visualization by Excel

Average_Website_Traffic	
985.625	
Date	**Website_Traffic**
2/29/2023	2500
44928	20
Correlation_Website_Sales	
0.224236537	

Create visual representations to illustrate findings:

- The average daily website traffic: 985.625
- The highest website traffic is 2500 on 2023-2-29
- The correlation between website traffic and sales: 0.224236537

4. Insights and Findings:

- The highest correlation of sales is the temperature, and the higher the temperature is, the higher the sales are.
- The highest sales happened in March, so marketing spending was also used most in this month.

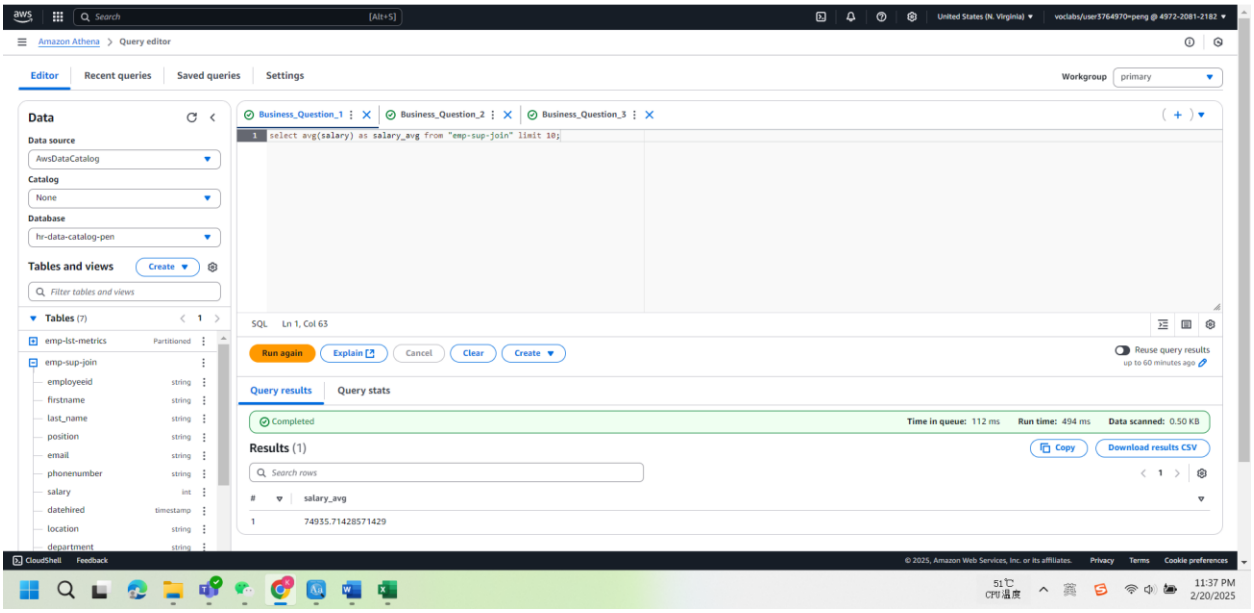
5. Recommendations:

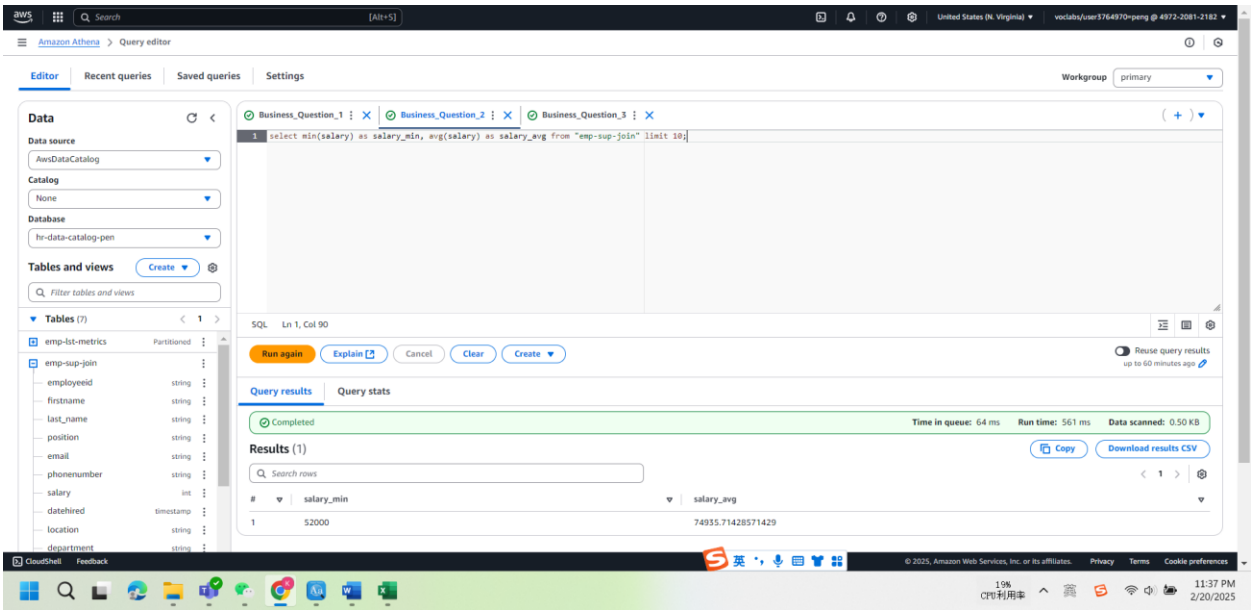
- The sales rely more on the temperature, so adjusting the suitable temperature in the store is essential, although the outside temperature cannot be controlled.

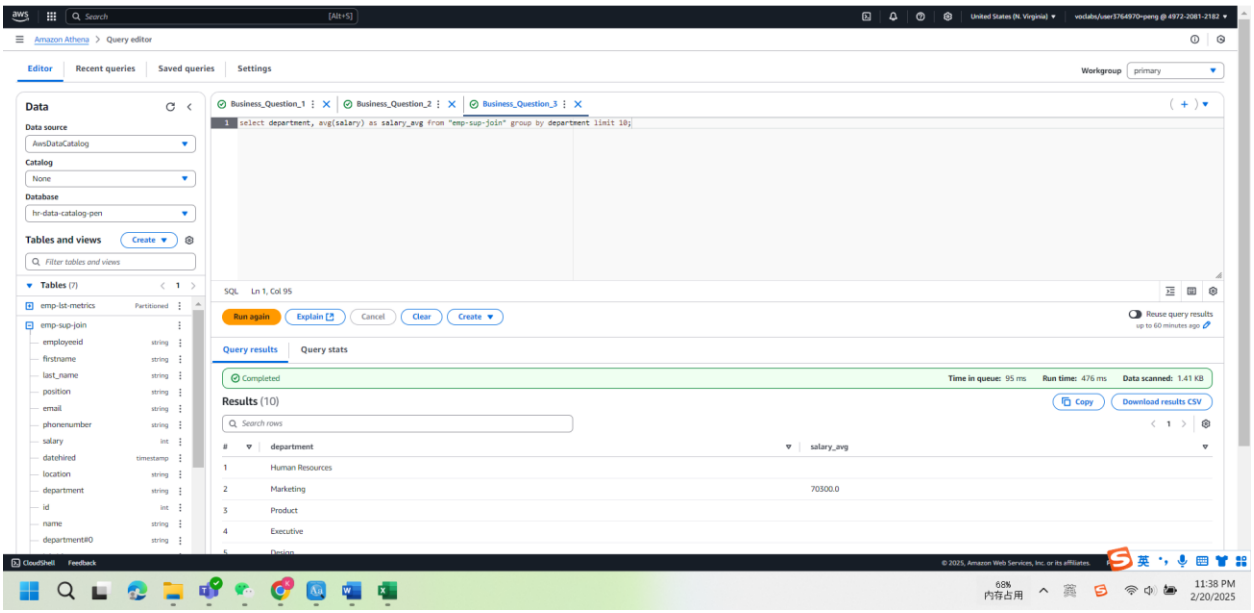
- The marketing expenditure can be reduced gradually, particularly in the month with higher sales.

6. Other Descriptive Analysis (Weekly Activity 6)

Figure 9. Using Athena to Do descriptive Analysis







Tools and Technologies:

- SQL and Excel for data analysis, and AWS Athena for descriptive data analysis
- Data visualization tool is Excel

Deliverables:

- A detailed report summarizing the methods, findings, and recommendations.
- Visualizations and dashboards to present key insights clearly.
- A presentation for stakeholders to communicate important findings and suggestions for future action.

This descriptive analysis project aims to comprehensively understand sales performance, the impact of marketing expenditures, the impact of the temperature on sales, and the website traffic insights.

Data Wrangling

Project Description: The City of Vancouver needs to migrate to AWS, and we implemented a data analytic platform.

Project Title: AWS Data Analytic Platform for The City of Vancouver

Objective: The primary goal of this project is to perform comprehensive data wrangling to prepare a robust dataset for the 311 Contact Center at The City of Vancouver. By cleaning, transforming, and consolidating data from various sources at AWS, the project aims to enhance the accuracy and usability of this data for subsequent analysis and reporting.

Background: The City of Vancouver needs to migrate its data to AWS, and I selected one of the datasets about the 311 Contact Center. However, this data needs cleaning, otherwise making it challenging to derive meaningful insights. Effective data wrangling will facilitate better decision-making.

Dataset: The data wrangling process will involve one dataset, including:

- Date: 2024-1-1 to 2025-1-31
- Calls_Offered: the time of 311 offering calls (seconds)
- Calls_Handled: the time of 311 handling calls (seconds)
- Calls_Abandoned: the time of 311 abandoning calls (seconds)
- Average_Speed_of_Answer: the time of average speed of answers from 311 (seconds)
- Service_Level: 311 service level (0-1)
- BI_ID: The ID number of providing services

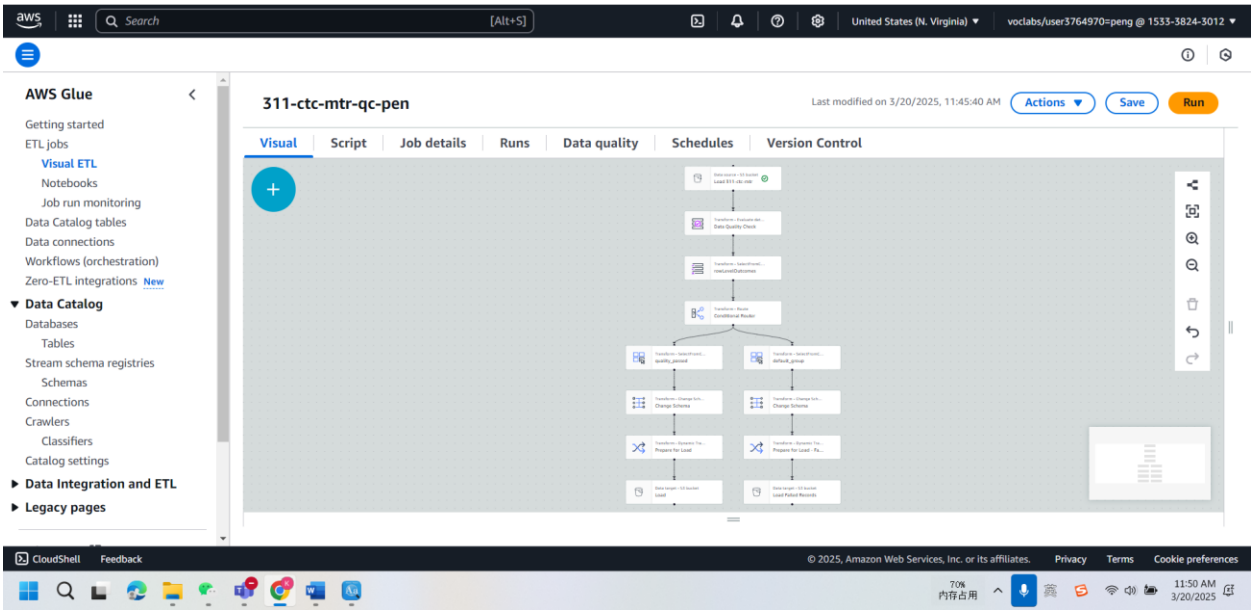
Methodology:

1. Data Collection:

- Download the dataset from the City of Vancouver
- Observing the initial dataset, which includes missing values, outliers, and data type.

2. Data Assessment and Cleaning:

Figure 10. Quality Check in Visual ETL



I did the ruleset for data quality and then set the conditional router, including passed and failed, and then changed the schema before load, including dropping the useless columns. Finally, I loaded the Passed and Failed into the corresponding folders.

3. Data Transformation and Consolidation:

Figure 11: Failed Dataset in Transfer Bucket

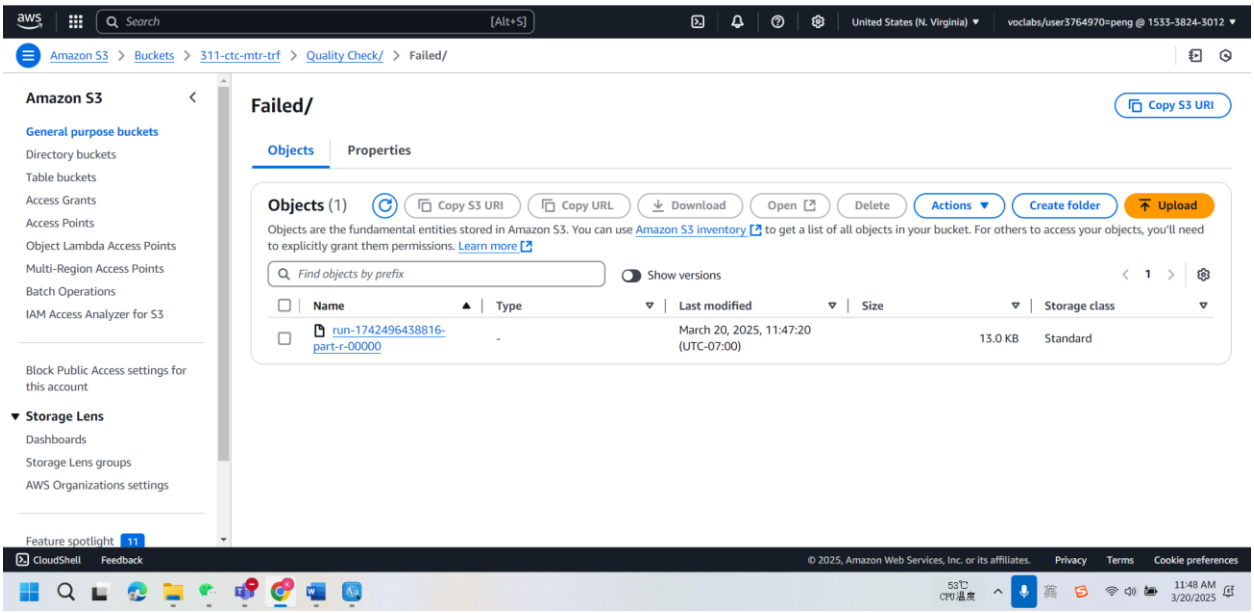
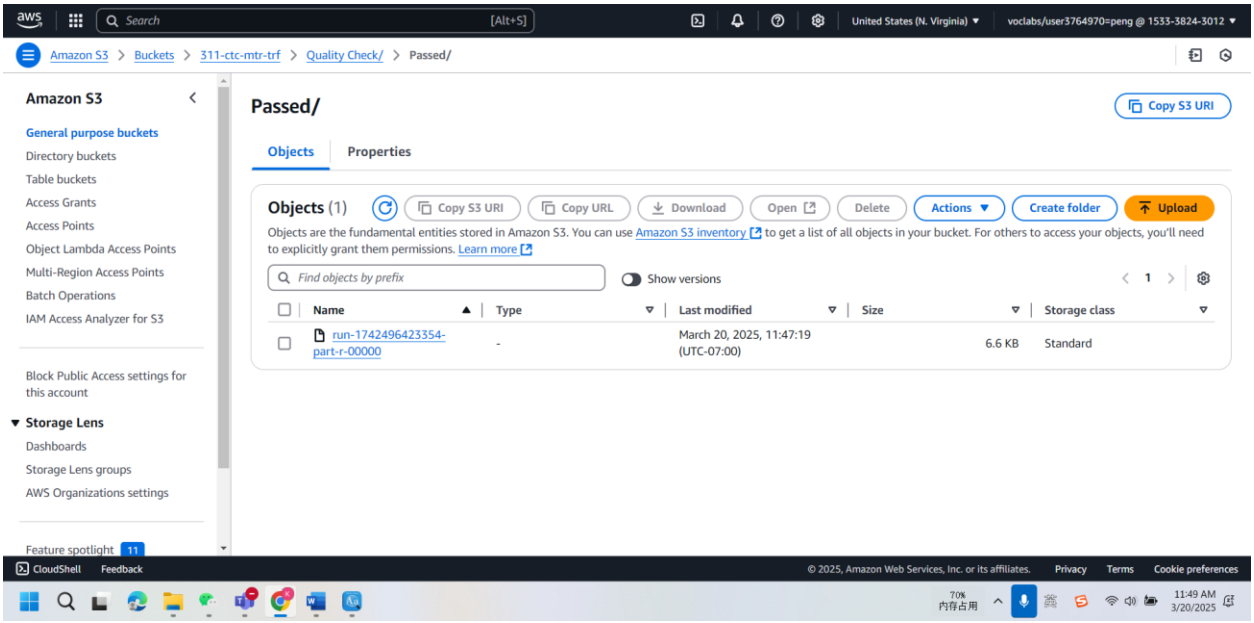


Figure 12: Passed Dataset in Transfer Bucket



After loading the passed and failed in Visual ETL, the corresponding datasets have already been published into each sub-bucket in the transfer bucket, which can be downloaded for different uses.

Tools and Technologies:

- AWS S3 and AWS Glue

Deliverables:

- A cleaned and transformed 311 Contact Center dataset ready for analysis, particularly for Passed Dataset in a CSV format.
- Confirmation of data quality checks conducted during the process.

Timeline:

- Expected completion of the project: Week 9 and Week 10, including phases for assessment, cleaning, transformation, and documentation.

This data wrangling project aims to establish a high-quality dataset that enables the City of Vancouver to conduct adequate 311 Contact Center Data.

Data Quality Control

Project Description: The HR Department of UCW implemented the Data Quality Control

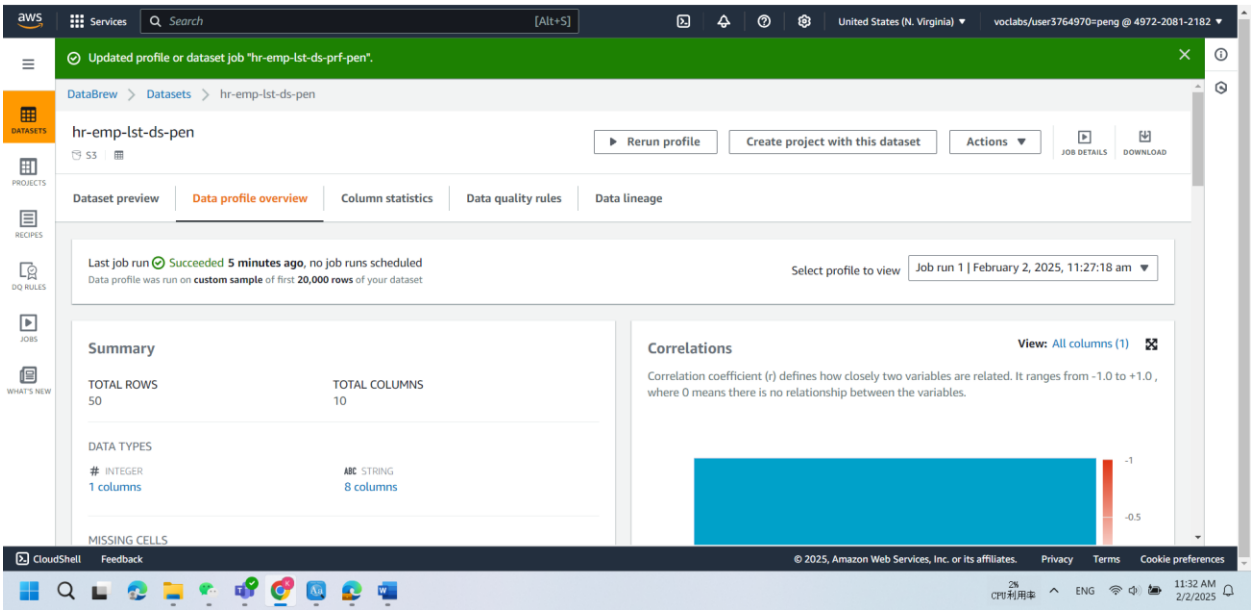
Project Title: Implementation of Data Quality Control Measures HR Department of UCW

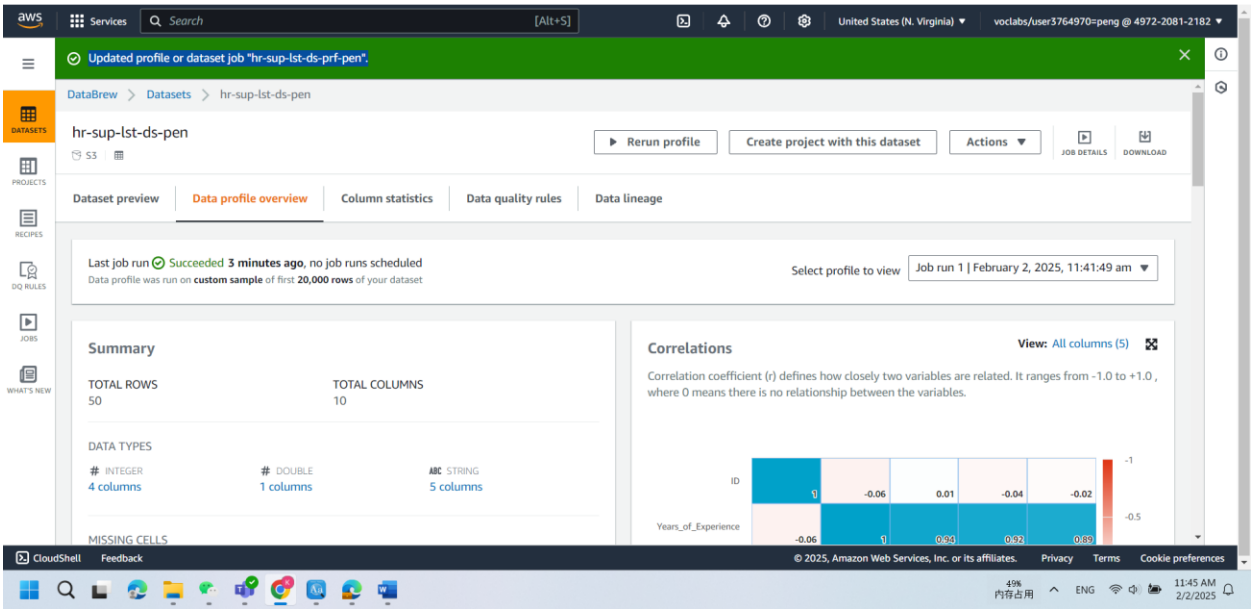
Objective: The primary objective of this project is to establish a comprehensive Data Quality Control (DQC) framework at AWS for the HR department of UCW. This framework will ensure the organization's data's accuracy, completeness, consistency, and reliability, enhancing decision-making processes and overall business performance.

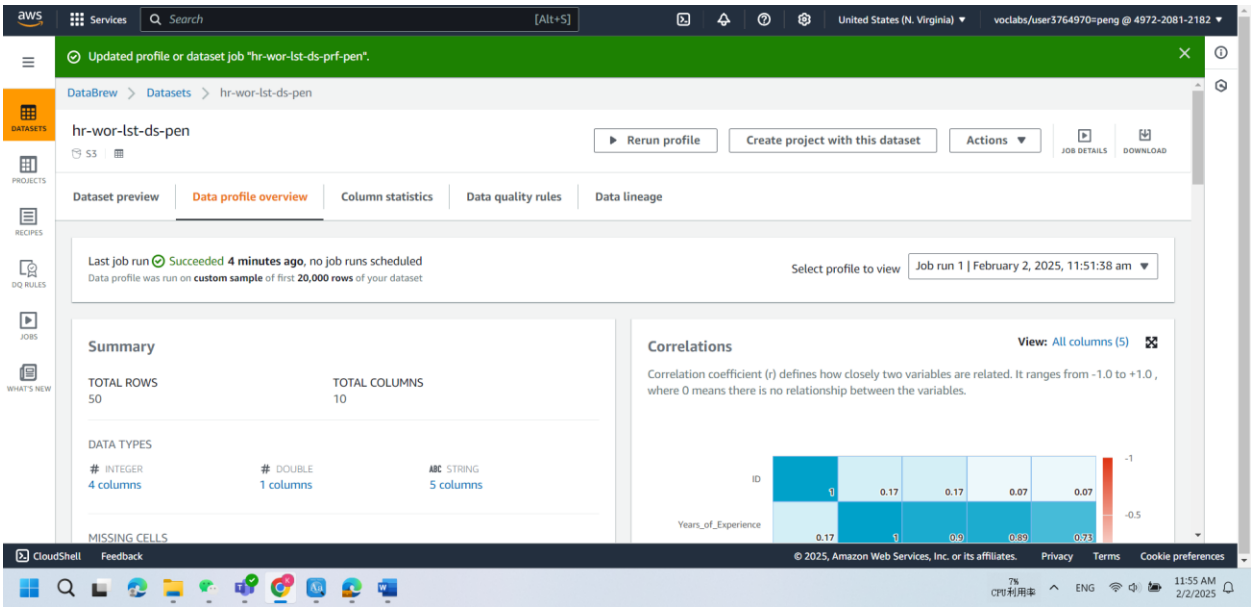
Scope: The project will focus on the following key areas:

- Data Profiling: Analyzing existing datasets to assess quality levels.

Figure 13: Data Profiling for HR of UCW

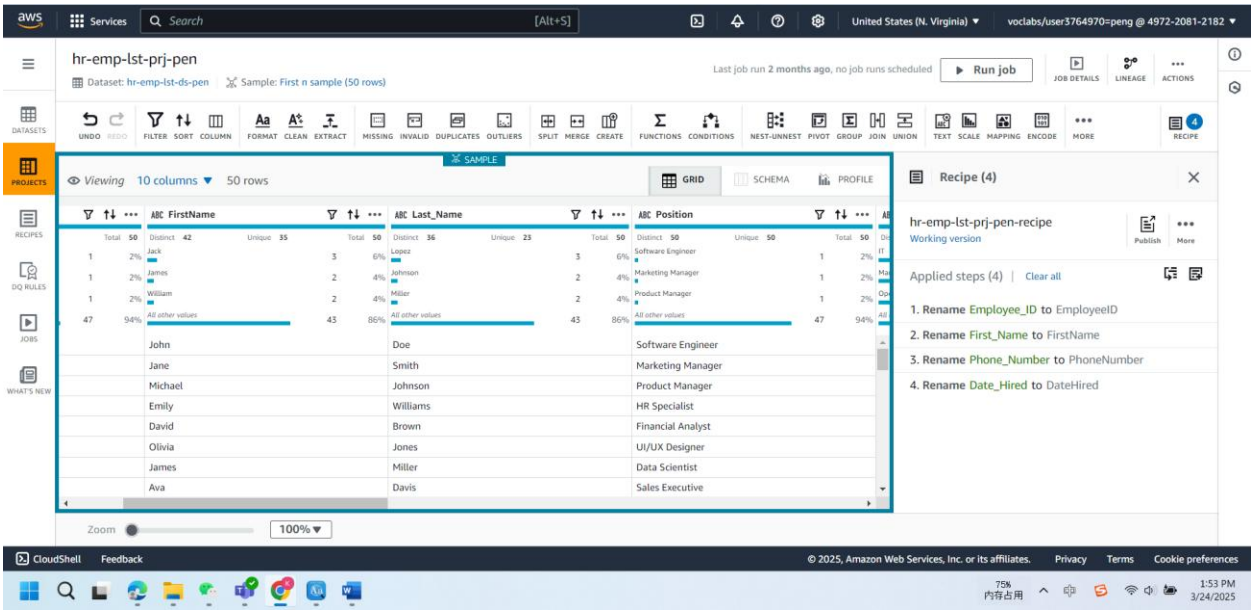


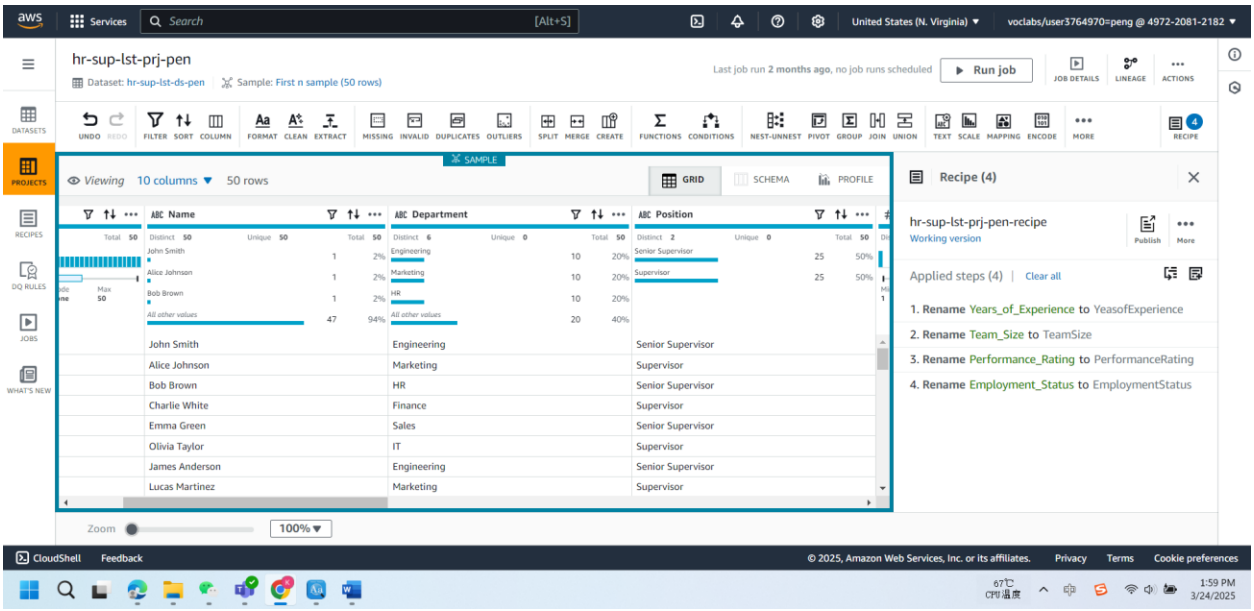


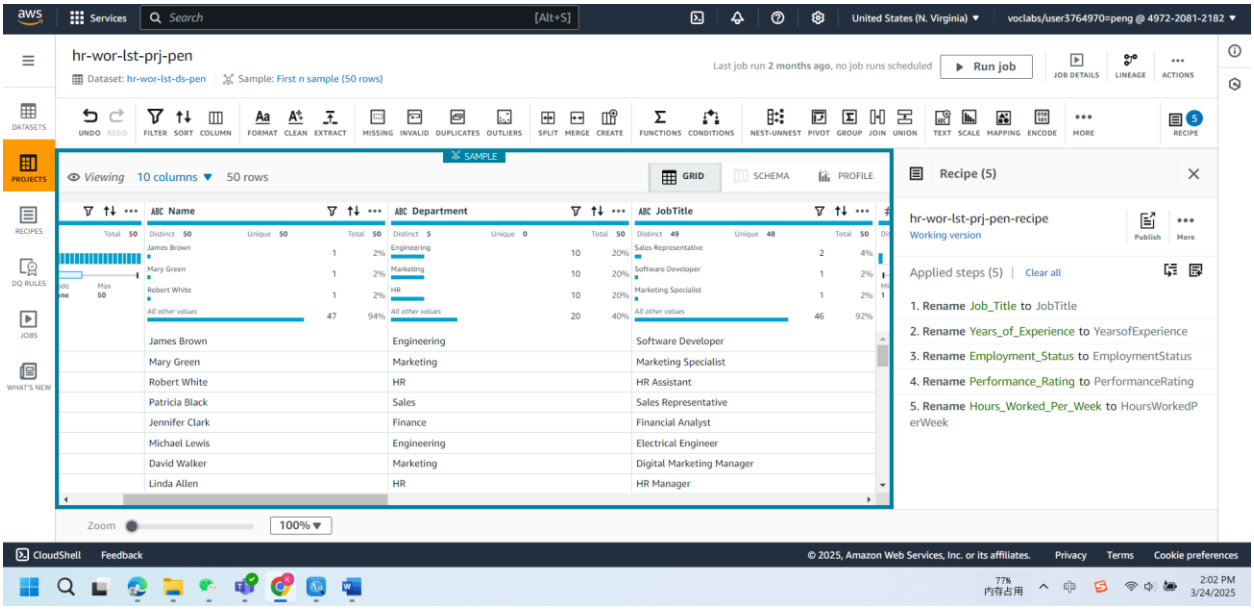


- Data Cleansing: Developing processes to correct inaccuracies and eliminate duplicates.

Figure 14: Data Cleaning for HR of UCW

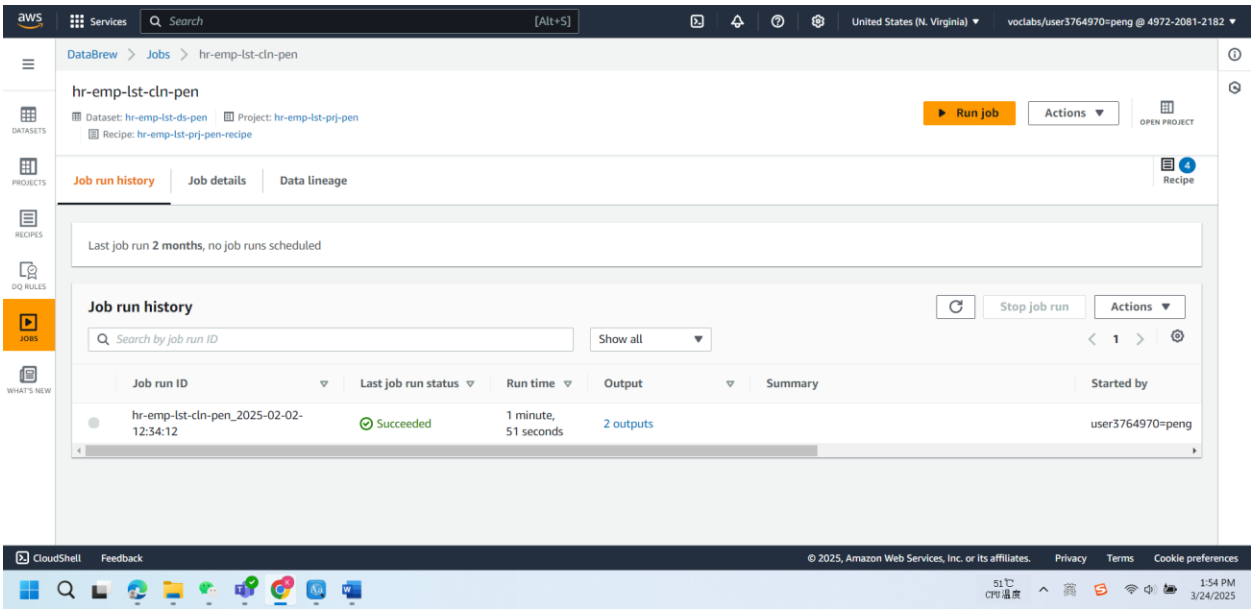


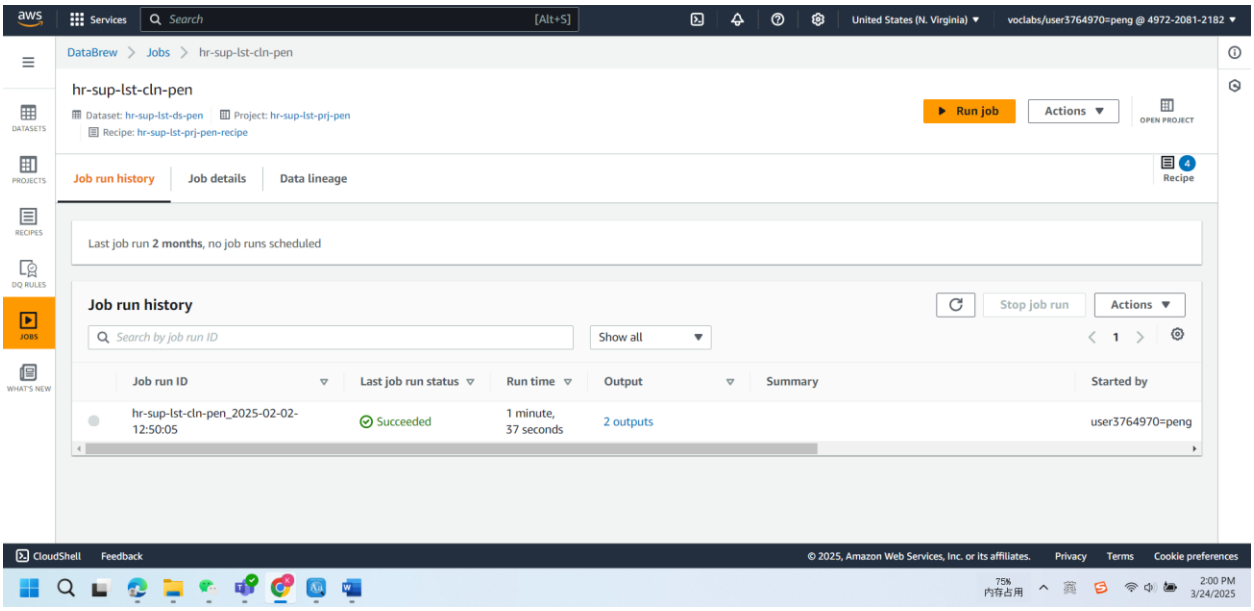


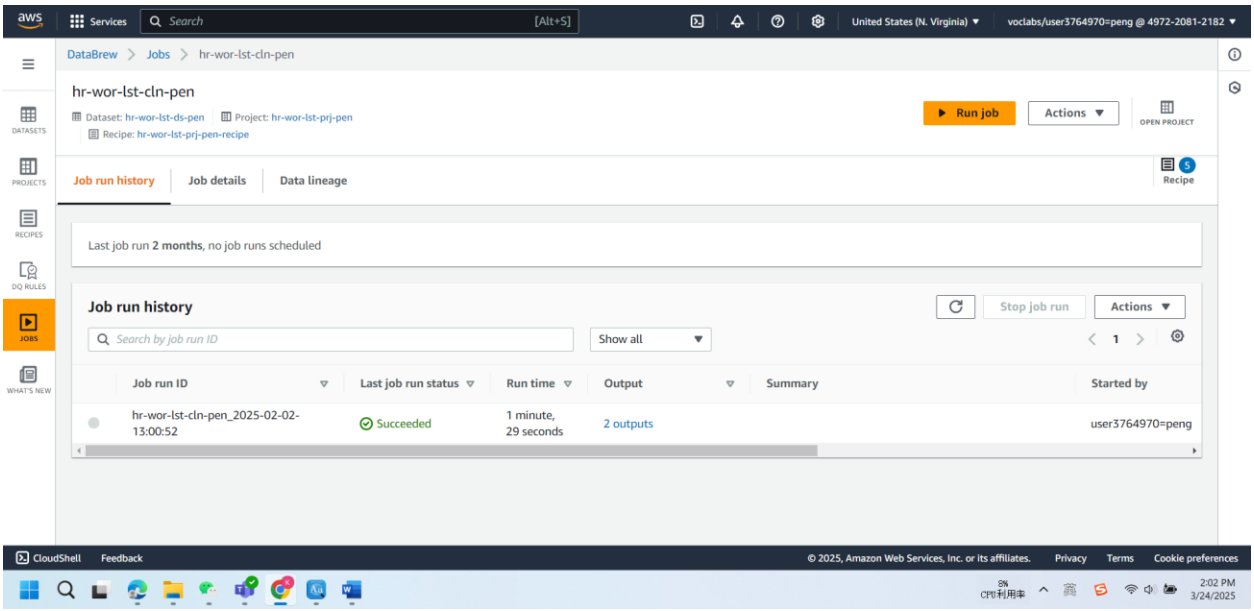


- Data Validation: Implementing validation rules and checks to ensure data integrity.

Figure 15: Data Validation for HR of UCW

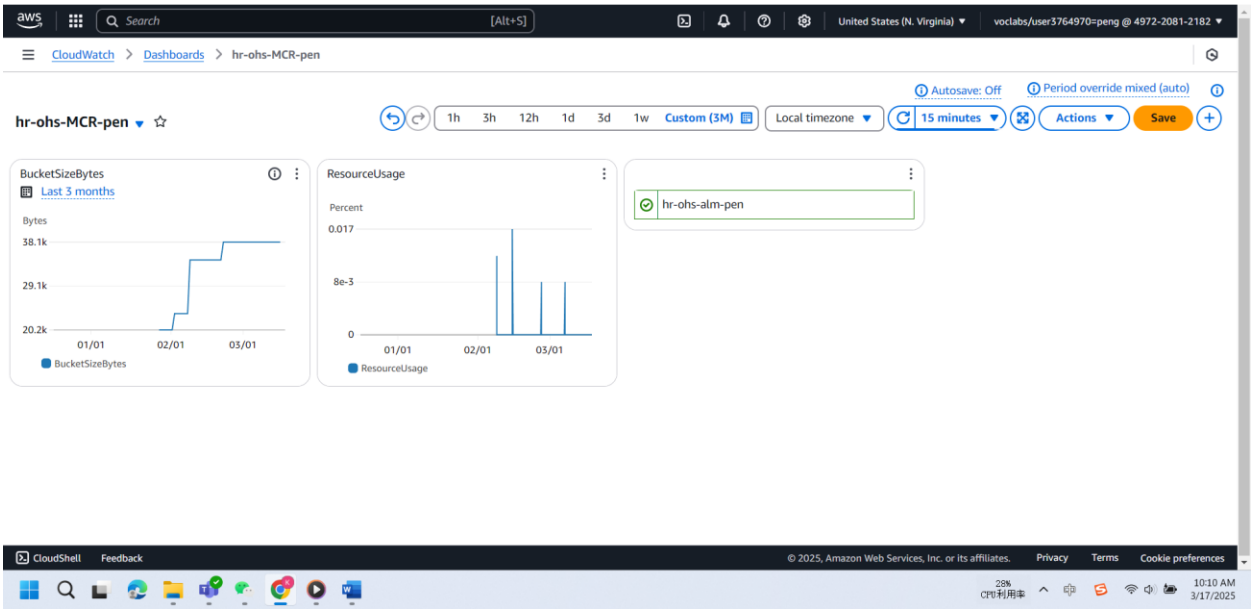

Services
Search
[Alt+S]
United States (N. Virginia)
voclabs/user3764970=peng @ 4972-2081-2182
DataBrew
Jobs
hr-emp-lst-cln-pen
hr-emp-lst-cln-pen
Dataset: hr-emp-lst-ds-pen
Project: hr-emp-lst-prj-pen
Recipe: hr-emp-lst-prj-pen-recipe
Run job
Actions
OPEN PROJECT
Job run history
Job details
Data lineage
Recipe
Last job run 2 months, no job runs scheduled
Job run history
Stop job run
Actions
Search by job run ID
Show all
Job run ID
Last job run status
Run time
Output
Summary
Started by
hr-emp-lst-cln-pen_2025-02-02-12:34:12
Succeeded
1 minute, 51 seconds
2 outputs
user3764970=peng
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences
1:54 PM
3/24/2025

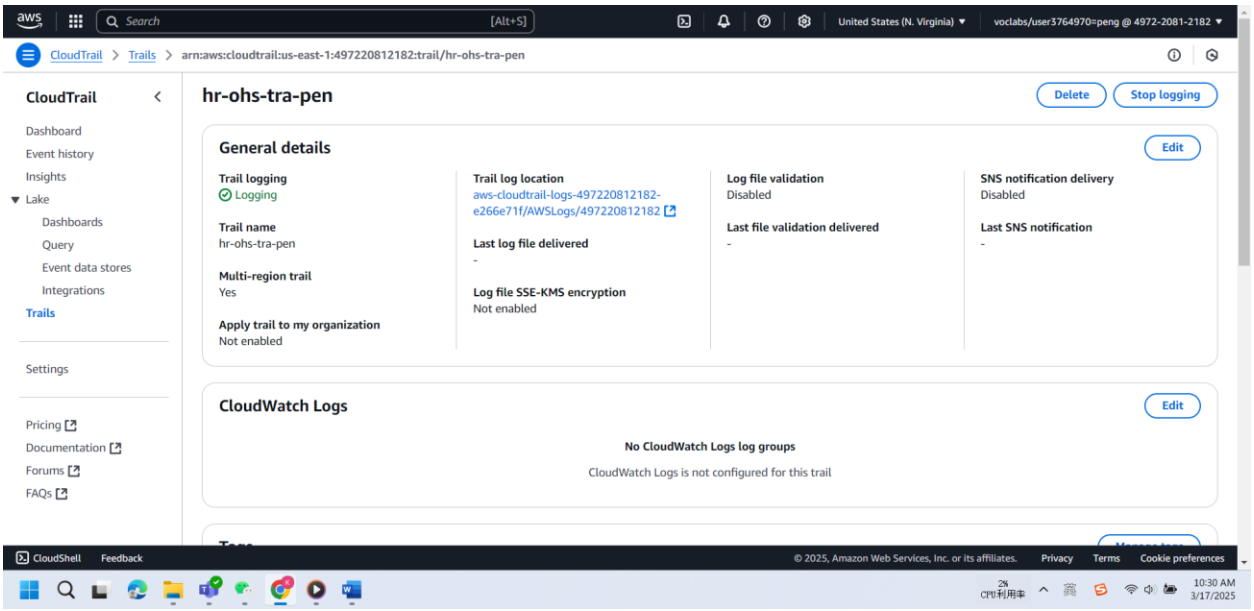

Services
Search
[Alt+S]
United States (N. Virginia)
voclabs/user3764970=peng @ 4972-2081-2182
DataBrew
Jobs
hr-sup-lst-cln-pen
hr-sup-lst-cln-pen
Dataset: hr-sup-lst-ds-pen
Project: hr-sup-lst-prj-pen
Recipe: hr-sup-lst-prj-pen-recipe
Run job
Actions
OPEN PROJECT
Job run history
Job details
Data lineage
Recipe
Last job run 2 months, no job runs scheduled
Job run history
Stop job run
Actions
Search by job run ID
Show all
Job run ID
Last job run status
Run time
Output
Summary
Started by
hr-sup-lst-cln-pen_2025-02-02-12:50:05
Succeeded
1 minute, 37 seconds
2 outputs
user3764970=peng
CloudShell
Feedback
© 2025, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences
2:00 PM
3/24/2025



- Monitoring and Reporting: Establishing ongoing monitoring processes and dashboards to track data quality metrics.

Figure 16: Data Monitoring for HR of UCW





Methodology:

1. Data Profiling:

- Utilize data profiling tools to assess the quality of identified datasets, focusing on completeness, uniqueness, validity, consistency, and accuracy.
- Document findings to highlight areas requiring immediate attention.

2. Data Cleansing Processes:

Develop and implement procedures for data cleansing, which includes renaming.

3. Validation Rules and Procedures:

- Set up validation rules for new data entries to reduce the risk of poor-quality data being introduced into the system.
- Create data entry guidelines to promote consistency and accuracy.

4. Monitoring:

- Implement monitoring tools and dashboards that provide real-time data quality metrics and alerts for significant deviations.

Tools and Technologies:

- AWS CloudWatch
- WAS Glue DataBrew

Deliverables:

- A comprehensive Data Quality Control plan detailing processes, metrics, and responsibilities.
- Cleaned and validated datasets ready for analysis and reporting.
- A monitoring dashboard that visualizes data quality metrics in real time.

Timeline:

Expected completion of the project: Week 3, Week 4, and Week 9.

This Data Quality Control initiative aims to empower UCW to enhance its HR data integrity and reliability, resulting in improved decision-making, operational efficiency, and compliance with regulatory requirements.